Homework for week 5 (due on Nov 27)

1. (35) For the dataset "BiologyClassificationWithTrueType.csv", use k-means to find classification and compare to the real types. In the dataset, there are 5 data fields, "X11", "X12", "X21", "X22", and "Type". The first 4 variables are attributes that you can use for clustering analysis. The last data column is the true category (you need to remove it before clustering analysis), and only used for comparison.
    a. (15) Build a k-means model on the dataset, and find the clusters in the data; justify your selection of K, the number of clusters.
    b. (10) Compare your results to the true category fields "Type", and draw your conclusions on K-means clustering analysis.
    c. (10) Describe any issues for the data, model and results.

2. (65) In this exercise, you will use the dataset "LapseData201410policy.csv". The variable names are quite self-explainable. The exposure is "ExposureN", and the lapse count is "LapseN". The lapse model has been built as "LapseN ~ offset(log(ExposureN)) + FaceAmount + PremiumMode + RiskClass + IssueAge + PremiumJump + I(PremiumJump^(-1)) + I(PremiumJump^(-2)) + I(PremiumJump^(-3)) + log(PremiumJump)+ PremiumMode:IssueAge". You, as a modeling actuary, are responsible to validate the model and generate plots to demonstrate the predictive power of the model.
    a. (20) Validate the model by using 60%/40% holdout method, for train/validation dataset, respectively.
    b. (20) Use 4-folder cross-validation method.
    c. (20) For both of above methods, generate an "Actual/Expected" table for each variable in the same format as shown in the class. Please refer to the class notes that was sent to you for the week 2, page 6.
    d. (5) Compare the two methods, and explain the difference.

    (Please note. Because the large number of records, the R algorithm may take up to a few minutes to finish, dependent on how fast your computer is.)

3. (Bonus 30) For the model in problem 2, try to improve the model performance by including higher order and/or cross terms. Prove that your new model have better model accuracy than the one specified in Problem 2 with some metrics that you can justify.